INSY662 – Final Individual Project

Tashfeen Ahmed – 261145667

Classification Model

Objective is to build a classification model predicting whether a project's state will be 'successful' or 'failure', based on the information available on a project's launch. Different classification models were compared, selecting best based on scoring metrics

1. Data Preparation: Valid predictors that are available at the launch of project and target variable are selected. Outliers are removed (Z-scores greater than 3) based on continuous variables ('name_len_clean', 'blurb_len_clean', 'converted_currency_USD'). Null values are also dropped. The 'goal' was converted to USD currency and a new column 'converted_currency_USD' is added, then 'goal' column along with 'static_usd_rate' is dropped.

Predictors					
goal	name_len	deadline_weekday	deadline_day	created_at_hr	create_to_launch_days
converted_currency_USD	name_len_clean	created_at_weekday	deadline_hr	launched_at_month	launch_to_deadline_days
static_usd_rate	blurb_len	launched_at_weekday	created_at_month	launched_at_day	
category	blurb_len_clean	deadline_month	created_at_day	launched_at_hr	

Numerical	Categorical	dropped	Column Created

Categorical Variables: The datetime variables (year, month, day, hour) are considered categorical. Day and hour are dropped in modeling process as their feature importance was low. Chi-square test could be used to eliminate categorical variable- however it is not taught. So, models were run repeatedly with and without a categorical variable to see if the model scoring metrics improve or not

Numerical Variables: Blurb_len and name_len were dropped due to high collinearity. Multi-collinearity can cause inaccurate results. Numeric Variables are Z-score standardized so each variable contributes equally to analysis. Moreover, Feature importance was done using random forest to identify important numerical variables

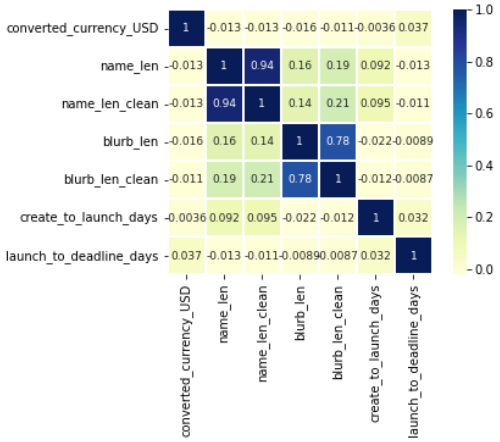


Figure 1: Correlation Heatmap for numeric Variables

2. Model Selection: GridSeachCV with 5-fold cross-validation was used to find best parameters for classification models such as number of trees for Random Forest, Gradient Boosting and hidden layer size for neural network. Dataset exhibited class imbalance for target variable, thus Macro F1-score was used as scoring parameter for choosing best parameters. It calculates metrics for each class independently and takes average – giving equal weight to each class.

3. Model Comparison: Following are the three results for 3 K-fold Validation (5 was taking too long to run):

Model	Fit Time	Score Time	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.15	0.01	69.10%	53.09%	42.52%
Random Forest	75.81	3.66	71.68%	59.47%	42.09%
Artificial Neural Network	48.33	0.02	64.29%	45.48%	46.18%
KNN Model	0.01	0.39	61.82%	39.40%	31.72%
Gradient Boosting	36.89	0.10	71.09%	56.32%	52.23%

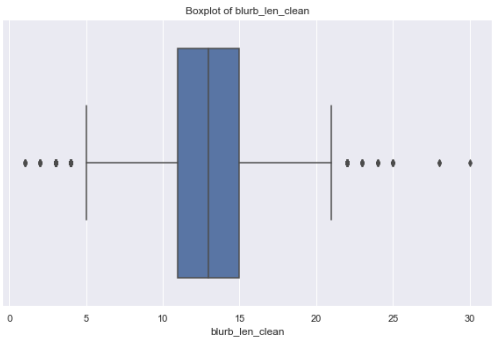
Gradient Boosting shows a superior balance, with a recall of 52.23% compared to Random Forest's 42.09%, indicating it's better at identifying true successes. Given the similar test accuracy and precision and a higher recall, Gradient Boosting is selected for the final model. Precision is more important when the consequence of predicting an unsuccessful project as a success is high. However, recall becomes more important when it's critical to identify all potential successful projects, even if that means including some false positives.

4. Test Set Preparation: Lastly code is written for testing on new data - all the preprocessing steps performed on original dataset are performed on it and all columns from original dataset are also added in the new dataset so that model trained on original dataset can be used.

Clustering Model

Goal is to group similar projects together. Most meaningful are considered.

1. Data and Methodology: Considering data had numerical and categorical variables all predictors were put on same scale using MinMaX Scaler. Continuous variables had large outliers found from Z-Statistics (Z-Scores greater than 3) and were removed. Categorical variable are dummified/encoded.



2. Results: K-means, KPrototype and DBSCAN are tried. K-means indicated 5 clusters with the best silhouette score of 0.357. There is some evidence of clusters present but not strong evidence (rule of thumb > 0.5). DBSCAN best parameters were found as eps=0.9 and min sample =3 using silhouette scoring metric. However, 509 clusters were formed – considering interpretability is difficult it is not preferred. KPrototype interpretation was difficult and finding optimal K was not easy

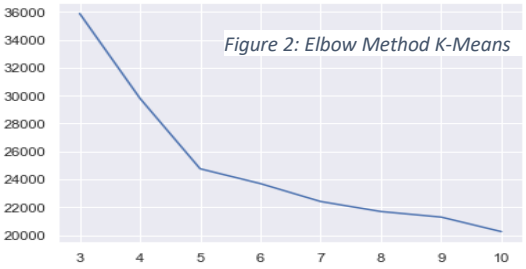


Figure 2: Elbow Method K-Means

as silhouette and elbow method cannot be directly applied with KPrototype which is something not taught in the course. With elbow method optimal K =5 was found for KPrototype

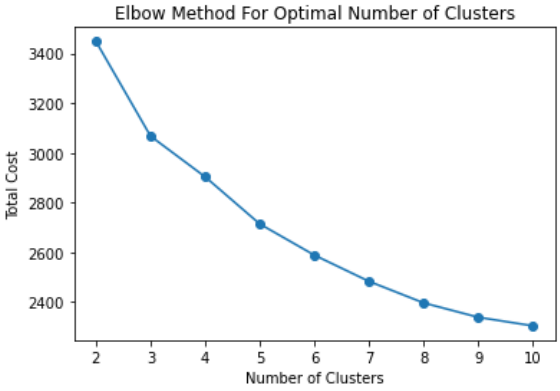


Figure 3: Elbow Method KPrototype

Numeric Variables: Overall, medium distinction between centroids. Cluster 1 and 3 have higher USD pledge, Clean name length, Blurb length clean than the rest. Create to launch day difference is almost similar for all clusters. Cluster 0, 2 and 4 have highest difference between launch to deadline days. Converted Currency USD which is the goal of a project Cluster 4,2 and 0 have higher average goals.

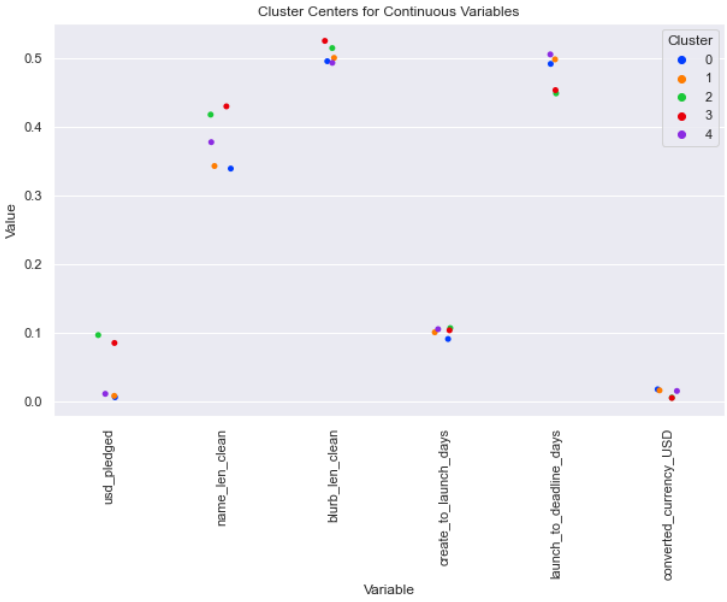


Figure 4: Cluster Center for Numeric Variables

Categorical Variables: Centroids indicate the average feature values for each cluster. In countries most startups are US Based relative frequency of ~ 81%, 66%, 68%, 81%, 64% for clusters 0, 1, 2, 3, 4 respectively. Cluster 0 and 3 had highest Staff pick at ~ 19% and 30%. Cluster 0 and 3 had highest Spoltilight at 100%. Cluster 0 and 3 have highest project success at ~100%. Cluster 2 and 3 are most launched and have deadlines in year 2014 at ~100% and 54% respectively. Cluster 0 and 1 are most launched and have deadlines in year 2015 at ~100%. In Cluster 0 categories such as Gadgets (13.5%) and Web(29.45%) are prevalent, Cluster 1 has a more proportion of Hardware(15.45%) and Web(25.32%), Cluster 2 includes Software (9.75%) and Hardware (30.9%), Cluster 3 includes Plays(15.87%) and Gadgets(14%), Cluster 4 includes Software(35.8%) and Hardware(43.37%)

The clusters suggest that successful projects tend to have a higher rate of staff picks, spotlight features, and are often in the Gadgets, Web and Plays categories. In contrast, failed projects often have lower staff picks and spotlight.

Please refer to Appendix for all cluster center values

Appendix

Cluster center Values

	usd_pledged	name_len_clean	blurb_len_clean	create_to_launch_days	launch_to_deadline_days	converted_currency_USD	state_failed	state_successful	country_AT	country_AU	country_BE	country_CA
0	0.005421	0.339065	0.495729	0.090413	0.491836	0.016947	1.000000e+00	-1.776357e-15	3.458367e-03	0.044427	3.192338e-03	0.062251
1	0.007434	0.342806	0.500609	0.100175	0.498396	0.015443	1.000000e+00	2.331468e-15	-1.973248e-17	0.051627	-1.691355e-17	0.072136
2	0.096136	0.417782	0.514898	0.106302	0.448832	0.004847	-1.332268e-15	1.000000e+00	-3.903128e-18	0.011472	2.233456e-17	0.031549
3	0.084601	0.429851	0.525407	0.103091	0.453517	0.004352	4.551914e-15	1.000000e+00	2.403846e-03	0.025240	6.009615e-04	0.038462
4	0.010399	0.377601	0.493390	0.104701	0.505642	0.014535	1.000000e+00	1.776357e-15	1.506024e-03	0.014307	3.765060e-03	0.030873

	country_CH	country_DE	country_DK	country_ES	country_FR	country_GB	country_IE	country_IT	country_LU	country_NL	country_NO	country_NZ	country_SE	country_US	staff_pick_False	staff_pick_True
0	4.256451e-03	2.420857e-02	0.010375	1.064113e-02	2.075020e-02	0.108540	0.006119	1.542964e-02	1.301043e-18	0.020218	0.004789	0.005853	0.006651	0.648843	0.977654	0.022346
1	3.903128e-17	1.387779e-16	0.004950	-5.030698e-17	1.734723e-17	0.144625	0.002475	1.613293e-16	1.544988e-18	0.026521	0.002475	0.008133	0.004243	0.682815	0.931047	0.068953
2	4.780115e-04	1.434034e-03	0.000956	-4.076600e-17	9.560229e-04	0.129063	0.001912	4.780115e-04	-1.179070e-18	0.004302	0.000478	0.003824	0.001434	0.811663	0.700765	0.299235
3	2.403846e-03	1.802885e-02	0.005409	3.004808e-03	1.502404e-02	0.192308	0.004808	1.201923e-03	6.009615e-04	0.015024	0.001803	0.004207	0.004207	0.665264	0.807091	0.192909
4	7.530120e-03	8.283133e-03	0.004518	5.271084e-03	9.789157e-03	0.070783	0.003765	1.204819e-02	-8.267042e-19	0.002259	0.003012	0.003012	0.003765	0.815512	0.963102	0.036898

	category_Academic	category_Apps	category_Blues	category_Experimental	category_Festivals	category_Flight	category_Gadgets	category_Hardware	category_Immersive	category_Makerspaces	category_Musical
0	2.660282e-03	0.058260	4.119968e-18	0.013035	0.016760	0.028199	0.134876	0.089119	0.011705	0.012769	0.036712
1	1.414427e-03	0.094767	3.794708e-18	0.014144	0.014851	0.023692	0.092999	0.154526	0.012023	0.005304	0.029703
2	-1.452831e-17	0.044455	2.868069e-03	0.026769	0.048279	0.013862	0.066444	0.308795	0.021989	0.011472	0.071224
3	-1.084202e-18	0.074519	2.403846e-03	0.046274	0.062500	0.013822	0.140625	0.117788	0.034856	0.016827	0.090745
4	7.530120e-04	0.015813	3.577867e-18	0.003012	0.007530	0.009036	0.026355	0.433735	0.003012	0.004518	0.009789

Compiling Category information for each cluster for easy interpretation